Basic Concepts and Principles of Social Experimentation

This is the second in a series of papers on the design, implementation, and analysis of social experiments. In the previous paper, we defined a social experiment as a comparison of the outcomes of two or more groups randomly assigned to different policy regimes. In this paper, we explore in more detail the considerations involved in constructing this comparison and in interpreting the resultant differences in outcomes. We discuss the basic concepts and principles involved in:

- Deciding whether to experiment;
- Specifying the experimental treatment;
- Specifying the outcomes of interest;
- Interpreting the treatment-control service differential; and,
- Interpreting treatment-control differences in outcomes.

Deciding Whether to Experiment

A social experiment generally begins either with interest in some new program or a desire to determine whether an existing program is achieving its objectives. Unfortunately, given the limited resources available for program evaluation, not all new ideas and existing programs can be evaluated experimentally; sponsoring agencies must choose among competing uses of their evaluation budgets. The choice of policies or programs to be evaluated—and within that set the ones to be evaluated experimentally—requires a careful assessment of the likely value of the information to be obtained through rigorous evaluation and the cost and feasibility of obtaining it. In this section, we discuss the questions that should be addressed in deciding which programs or policy issues to investigate experimentally.

What does society stand to gain from the experiment?

A social experiment benefits society by providing better information on which to base public policy. Such information can improve policy in one of two ways: It can lead policymakers to adopt a program or policy that is found to have net social benefits (*i.e.*, benefits to society that outweigh its costs) or it can lead to the termination of an existing program that is found to have net social costs (*i.e.*, costs that outweigh its benefits to society).[1,2]

Of course, one cannot know before the fact whether any particular experiment will lead to a change in policy—that depends on the experimental findings and whether policymakers act on those findings. In deciding whether to conduct the experiment, then, one must act on the **expected value** of the experiment. This can be expressed as:

Expected value of experiment =
(Value of change in policy
× *Probability of change in policy)*
– *Cost of experiment*

For a new program, the value of a change in policy due to the experiment is the net social benefit that will accrue if the experimental program is adopted. For an existing program, the value of a change in policy is its net social cost; this is a measure of the resource savings that will accrue if the program is terminated. These values can be quite large, as witness the policy impacts of the National JTPA Study. That evaluation found that the out-of-school youth component of JTPA had essentially no impact on the earnings of

[1] The following line of argument is stated more formally in Burtless and Orr (1986). We will discuss in a subsequent paper how net social benefits and costs are estimated.

[2] In principle, experiments could also *prevent* the termination of an *effective* existing program or *prevent* the adoption of an *ineffective* new program. The former is analytically identical to the case in which the experiment leads to the adoption of an effective program and the latter is identical to the case in which the experiment leads to the termination of an ineffective program.

youths who participated in it. As a direct result of this finding, funding for the out-of-school youth component of JTPA was reduced by over $500 million per year. In just a few years, this savings of resources that would otherwise have been wasted on ineffective training services easily surpassed the cost of all the social experiments that have been conducted in the last 30 years.[3,4]

It is obviously difficult to predict either the value of a change in policy or the probability it will occur, but one can make some statements about them that are useful in discriminating among potential experiments. First, other things equal, the larger the program the larger its social benefit or cost is likely to be. Thus, social experiments focused on larger programs are likely to have higher social value. Second, in some cases previous research may allow one to make at least qualitative statements about the probability that a new program will be found to be effective or an existing program to be ineffective. The more credible nonexperimental evidence there is that a new program may be effective or that an existing program may be ineffective—*i.e.*, that an experiment would indicate that a change in policy is warranted—the higher the expected value of the experiment is likely to be.[5]

The social value of an experiment depends not only on the inherent importance and validity of the information it provides, but also on whether it is used to improve policy. An experiment is of no value to society if its results never influence policy. It is, of course, extremely difficult to predict *a priori* whether a particular set of evaluation results will be acted upon. Evaluation is only one of a number of forces impinging on the political process and in many cases not the most important one. Still, one can identify certain factors that make it more or less likely that evaluation results will play a key role in policy deliberations. For example, the results are more likely to influence policy if the behavioral questions that evaluation can address are central to the policy debate than if policy decisions turn on philosophical or ideological issues. To cite two extreme examples, job training programs are based almost entirely on the premise that the services they provide will increase the employment and earnings of participants, a behavioral premise that can readily be tested experimentally, whereas Social Security benefits for the aged are justified primarily on equity grounds, without regard to any behavioral effects they may have.

The likelihood that evaluation results will be acted upon will also be influenced by their timing, relative to the life-span of the policy issue they address. Social experiments take time to plan, implement, and analyze. Often the treatment itself lasts a year or more and several more years may be required to observe the outcomes of interest and analyze the program's impact on them. The interval between the decision to mount an experiment and the availability of results is often five to ten years. Only if an experiment addresses a relatively fundamental policy issue will its results still be relevant to the policy process after such a lag.

The income maintenance experiments and the health insurance experiment are examples of social experiments that focused on fundamental policy issues that were still relevant many years after the experiments were completed. Rather than estimating the impact of a specific policy, these experiments were designed to estimate underlying behavioral parameters—the elasticity of supply of labor and the price elasticity of demand for medical care—that would be relevant to a wide range of policies. And while these experiments never resulted in the adoption of any specific income maintenance or health insurance programs, their results have been used extensively in the analysis of a number of proposed programs and policies in these areas. Evaluations of ongoing programs are also highly likely still to be relevant when their results become available, because the program is likely still to be in place.

In contrast, novel program or policy proposals may have such a short life-span that they are irrelevant to policy by the time an experimental test can be conducted. This is particularly true if the proposal has only limited support to begin with—*e.g.*, a proposal developed by a single government official without significant support in the rest of the executive branch or the legislature. Neither that official nor the proposal are likely to be around five years later.

Proposals that involve a complex package of programmatic components are particularly susceptible to shifts in policy interest away from the specific combination of program elements evaluated before the results become available, even though there may still be substantial interest in its individual components. The experimental evaluations of state welfare reform demonstrations conducted in recent years illustrate the problem of experimenting with complex pro-

---

[3] See Greenberg and Shroder (1997) for a catalog of the social experiments that have been conducted and their cost.

[4] This result benefited not only the taxpayers, but also the disadvantaged youths who were the intended beneficiaries of the program. Rather than perpetuating a program that wasted their time and raised false expectations, the government initiated a search for more effective ways to improve youths' earnings. Whether that search will be successful depends on the outcome of several experimental tests of youth training programs that are underway as this is written.

[5] This presumes, of course, that the nonexperimental evidence is not sufficiently compelling to convince policymakers to make the change in policy without the benefit of an experiment. This may often be the case, however, because of the inherent risk that nonexperimental evidence may be contaminated by selection bias or the other threats to validity discussed in the first paper in this series.

grams. Many of these demonstrations involved multiple policy interventions intended to reduce the dependence and increase the self-sufficiency of welfare recipients—*e.g.*, employment and training services, child care assistance, enhanced medical care, financial incentives to work, time limits on receipt of assistance, and elimination of benefit increases for additional children.[6] A demonstration evaluation designed only to estimate the impacts of the overall policy package will have only very limited policy applicability; strictly speaking, its results apply only to that specific policy package. A much more powerful, versatile approach is to measure the impacts of the individual program components and/or alternative combinations of closely related components, so that the impacts of other policy packages can be inferred. In a subsequent paper, we will examine how experiments can be designed to do this.

What would it cost to conduct an experiment?

Against the potential value of an experiment must be weighed its expected costs. These include the costs of project planning, implementation and monitoring of random assignment, data collection, and analysis. The extent to which the costs of the experimental program itself represent a net cost to society depends on whether the experimental services generate social benefits—which in most cases cannot be known until the experiment has been conducted.[7] For planning purposes, it is probably prudent to treat these services, or some proportion of them, as a cost.

The costs of alternative experiments can differ enormously, depending on the sample sizes required to measure impacts with adequate precision and the method, frequency, and duration of data collection.[8] A typical social experiment costs $2-3 million, although it is quite possible to conduct one for substantially less and some, such as the Seattle-Denver Income Maintenance Experiment and the Health Insurance Experiment, which involved intensive, long-term data collection, cost over $80 million. In choosing among potential experiments, then, it will be important to obtain accurate estimates of the costs of each. Fortunately, once a design has been specified, it is possible to predict the costs of an experiment fairly accurately.

What are the alternative sources of evaluation information?

The benefits and costs of social experiments must be judged relative to those of the next best alternative source of information. If a reliable nonexperimental evaluation already exists, or could be conducted at little cost, an experiment may not add sufficient information to justify its cost. In a subsequent paper, we will discuss in more detail the relative strengths and weaknesses of experimental and nonexperimental analyses.

In deciding whether to rely upon nonexperimental evidence, it is important to bear in mind the inherent risk of nonexperimental methods: unlike experimental estimates, one can never be sure that nonexperimental estimates are unbiased. One must therefore examine carefully, and be prepared to accept, the assumptions on which any nonexperimental estimates are based. One should also apply several different nonexperimental methods to see if they yield similar estimates, rather than simply accepting the results of a single method. Recall that it was the implausibility of the assumptions required to estimate the labor supply response to transfer payments nonexperimentally that led to the income maintenance experiments and the inconsistency of nonexperimental estimates of the impacts of training that led to the National JTPA Study.

Is an experiment ethically and conceptually feasible?

In the first paper in this series, we discussed the ethical considerations involved in conducting an experiment. It is important that each prospective experiment be reviewed carefully with respect to these considerations to ensure that it can ethically be undertaken. As noted in that discussion, in some cases it may be possible to make an otherwise unethical experiment acceptable by changing the design somewhat or compensating the participants.

Potential experiments should also be reviewed for their conceptual feasibility. Some policy interventions are inherently inconsistent with random assignment at the individual level. For example, it is impossible to insulate a control group from the effects of a public education campaign conducted through the general media, or one that attempts to change the educational philosophy of an entire school system. While, as we will discuss in a later paper, it is conceptually possible to evaluate such interventions by randomly

---

[6] The 70 welfare reform demonstrations approved over the period 1989-96 averaged approximately seven distinct policy interventions per demonstration, according to Wiseman (1996).

[7] In the special case where the experimental benefits take the form of cash or near-cash transfers (e.g., food stamps or housing subsidies), it can be assumed that the benefits to transfer recipients equal the cost to taxpayers, so that the net cost to society is zero.

[8] These design issues will be discussed in subsequent papers.

assigning groups of individuals, such as entire communities or school systems, that approach has severe limitations in many contexts.

Specifying the Experimental Treatment

The **experimental treatment** is the offer of services to, or the imposition of policies upon, the treatment group that are not offered to or imposed upon the control group. The treatment is usually synonymous with the program or policy being evaluated or considered for adoption on an ongoing basis. For example, in the National JTPA Study, the treatment group was offered entrance to JTPA while the control group was barred from the program. In testing a new program, one would generally try to replicate as exactly as possible the features one would expect if it were adopted as an ongoing program.

This means that in the case of a new program, specification of the treatment involves codification of a set of rules and procedures as detailed as the statutes and operating procedures that govern a regular program. In social experiments where the treatment is administered by the research team, these procedures must be developed *de novo*; this can be a daunting task. The researchers who designed the income maintenance experiments, for example, developed rules for "countable" income, deductions from income, reporting requirements, filing unit composition, accounting periods, and appeals procedures as complex as those embodied in the Internal Revenue Code.[9] The Health Insurance Experiment rules and procedures combined the complexity of a comprehensive health insurance policy with the income accounting and reporting rules required to administer income-conditioned insurance provisions.

Increasingly, experimental tests of new programs have relied on existing administrative agencies to deliver the treatment. For example, most of the welfare reform experiments of the 1980s and 1990s were carried out by local welfare agencies. This has the advantage not only of relieving the researchers of developing a voluminous procedures manual, but also ensures that the program is administered more like an ongoing program would be. Reliance on existing administrative agencies should not, however, relieve the researchers of the responsibility of examining the practices, procedures, and philosophy of those agencies to ensure that they are consistent with the program being tested. For example, in one experimental test of employment and training services for women, it was belatedly discovered that one of the service providers routinely counseled women not to accept employment because they would lose their welfare grants.

It is important to recognize that the experimental treatment is defined as the *offer* of the program or the *imposition* of policy, not the actual receipt of program services or compliance with the experimental policy. It is the offer of service or the imposition of policy that automatically follows from random assignment and therefore definitively distinguishes the treatment group from the control group. The actual receipt of program services or compliance with the experimental policy is an experimental outcome that may or may not occur. The importance of this distinction is that the difference in outcomes between the treatment and control groups reflects the response of the entire treatment group to the offer of services or imposition of policy, whether they actually received those services or complied with the policy. We discuss below how the impact of the *receipt* of services on program participants can sometimes be inferred from the impact of the offer of services on the entire treatment group. This distinction also has important implications for the design of experiments, to be considered in a subsequent paper.

As noted above, the treatment is usually synonymous with the program or policy that is being considered for adoption on an ongoing basis. In certain instances, however, this is not the case. In the Manhattan Bail Bond Experiment, for example, the policy of interest was pretrial release without bail.[10] However, the researchers did not feel that they could persuade judges to agree to automatically release defendants without bail on the basis of random assignment to the treatment group. Therefore, the treatment in this experiment was a *recommendation* to the judge that the defendant be released without bail. Fortunately, the judges accepted a high enough proportion of these recommendations to produce a meaningful difference in release rates between the treatment and control groups. Nevertheless, in interpreting the results of this experiment, it must be borne in mind that not all of the members of the treatment group were released before trial.

Experiments designed to estimate behavioral responses, rather than to test specific programs, are also cases in which the treatment may differ from the one that might be adopted on an ongoing basis. The Health Insurance Experiment,

[9] See Kershaw and Fair (1976) for a detailed description of the administrative procedures developed for the New Jersey Experiment.

[10] See Botein (1965).

for example, was designed to estimate the price elasticity of demand for a broad range of medical services.[11] To achieve this objective, the researchers deliberately designed the experimental insurance policies to include a much broader scope of benefits than was likely to be included in any governmental program. The cost-sharing provisions (deductibles and coinsurance) in the experimental policies were also much simpler than those likely to be included in a government program, in order to allow direct estimation of demand elasticities. The intent of the study was that these elasticities could then be used to estimate the utilization of medical care under a wide range of health insurance policies.

Specifying the Outcomes of Interest

The fact that, at the point of random assignment, the treatment and control groups do not differ systematically in any way except eligibility for the experimental treatment means that *any* subsequent systematic difference in outcomes can be confidently attributed to the program. (By "outcome", we mean any behavior or events that occur after random assignment; we discuss below what we mean by "systematic".) Only a limited number of outcomes can be measured, however, because data collection is costly. Thus, great care must be taken in choosing the outcomes to be measured. Three types of outcome data are usually collected in social experiments—those related to program participation, achievement of program objectives, and other benefits and costs of the experimental program.

Program Participation

As noted above, the experimental treatment is the offer of services or the imposition of policy. It is important to measure the extent to which program services were actually received or the treatment group members complied with the experimental policy. This information will be critical in interpreting the impact estimates. It is sometimes the case, for example, that experimental programs had little or no impact because the services were not delivered as intended or because the treatment group did not comply with the experimental policy. More generally, it is important to document the services provided by the experiment so that policymakers know what intervention produced the estimated impacts.

Documentation of services received may also help suggest ways to deliver program services more efficiently or increase compliance with the experimental policy, and it may be helpful in planning for the implementation of the program or policy on an ongoing basis. A final use of data on program participation is in the procedure described below for inferring impacts on program participants when some treatment group members do not participate in the experimental program. This procedure requires individual-level data on service receipt.

The type of program participation data to be collected will obviously vary with the type of program being tested. In general, however, it should include data on whether and when each sample member entered and left the program, the amount and type of services (or other benefits) received, and when they were received, as well as narrative descriptions of the nature of the services and the service provider.

When services similar to those provided by the experiment are available from nonexperimental sources, it is also important to document receipt of nonexperimental services by both the treatment and control group. As we shall see below, the impact of the program will be determined by treatment-control differences in the combination of experimental and nonexperimental services.[12] Again, knowing the services received by the two groups can be critical to explaining the impact estimates. Suppose, for example, that an experimental employment program is found to have no impact, but the participation data show that the experimental services simply substituted for similar services that the treatment group would have received from other sources, such as the Employment Service. We would conclude that the effectiveness of the experimental services had not really been tested, because the experiment failed to create a treatment-control difference in total services. (We discuss in more detail below the interpretation of the impact estimates when services similar to those provided by the experiment are available elsewhere.)

Achievement of Program Objectives

Social programs are intended to address some problem afflicting individuals. The objectives of the program can usually be stated with reference to that problem. For example, training programs are intended to deal with the problems of unemployment and low earnings due to low skills. Their objectives, therefore, are to increase the employment and earnings of their participants. Prenatal nutrition programs are intended to address the problem of

[11] See Newhouse (1993).

[12] This does not mean that we assume that experimental and nonexperimental services are equally effective. It simply means that any treatment-control difference in nonexperimental services can lead to differences in other outcomes, just as a treatment-control difference in experimental services can. Thus, both must be measured.

poor diet among low-income pregnant women, which frequently results in unhealthy babies. Their objectives are to improve the nutrition of expectant mothers and, therefore, the health of their babies.

To measure whether a program is achieving its stated objectives, an evaluation must define those objectives in terms of measurable outcomes, such as employment and earnings, the nutrition of expectant mothers, or the birth weight of infants. These are the outcomes on which program impacts will be estimated. As these examples suggest, programs may have multiple objectives. In order to provide a comprehensive evaluation of the program, it is important that the evaluation identify and specify measurable outcomes corresponding to as many of the program's objectives as possible. In doing so, it is essential that the researchers consult with policymakers and practitioners in the field, both to gain their insights with regard to program objectives and to avoid the possibility that, after the experimental analysis has been completed, practitioners will point out some critical omission in the impacts measured by the experiment.

In some cases, it is useful to measure intermediate, as well as final, program objectives, to elucidate the mechanisms through which the intervention has its effects, or to see why it failed to have an effect. Consider, for example, an experimental parenting program that is intended, among other things, to improve children's school performance. In such an experiment, it would be possible to measure the extent to which the parent understands the principles taught in the program, the extent to which she applies them, the extent to which they change her children's behavior, and the extent to which their altered behavior affects their school performance. Failure of the program to affect school performance could be the result of a breakdown of any of these linkages.

Remarkably enough, policymakers and program administrators sometimes cannot agree on the objectives of the programs they administer, or they may view delivery of the service as an end in itself. For example, some may view a child care program as aimed primarily at allowing mothers to work, others may see its objective as improving the child's social and cognitive skills, while yet others may view the delivery of "quality" child care as an end in itself. The evaluation need not resolve such disagreements, although prior discussion of program objectives is sometimes helpful in fostering agreement among those involved in the program. The evaluation's job is simply to measure all of the outcomes that may be viewed as objectives of the program. Once program impacts on those outcomes have been estimated, it will be up to the political process to decide whether those impacts justify the cost of the program.

Other Benefits and Costs of the Experimental Program

In designing a social experiment, it is important to try to anticipate *all* benefits and costs of the program, not just those that are directly related to the program's objectives. For example, interventions like education and training or community service programs require a substantial investment of time on the part of their participants and may therefore divert them from employment in the regular labor market. Thus, one impact of such programs may be to reduce the earnings of participants while they are in the program. It is important to measure such forgone earnings, as they may be an important cost of the program.[13]

It is, of course, important to collect data on the cost of the experimental program itself, including any benefits to participants. As with the behavioral impacts of experimental programs, the guiding principle for measuring program costs is to include all costs, and only those costs, that would not have been incurred in the absence of the program.[14] As this implies, program costs are best measured as experimental impacts. For example, suppose that an educational program for welfare mothers causes them to stay on assistance longer than they otherwise would have. The cost of additional welfare benefits to participants can be measured as the difference in mean benefits between the treatment and control groups. Similarly, any savings in welfare benefits could be measured as treatment-control differences.

The responsibility for anticipating program impacts that are not directly related to program objectives—especially adverse impacts—nearly always falls entirely on the researchers designing the experiment. Policymakers and program managers tend to be advocates of the program and, therefore, to think only in terms of positive impacts. For example, in the design of one experiment, lengthy discussions with the managers of a set of community service programs generated a list of over 20 outcomes representing objectives of the programs; the forgone earnings of participants during the 9 to 12 months they were in the program were never mentioned as a possible effect of the programs.

Formal specification of a comprehensive benefit-cost framework *before* data collection plans are finalized is an essential step in ensuring that no important benefits or costs are

[13] In a subsequent paper, we will discuss measurement of the social benefits and costs of the program in a formal benefit-cost framework.

[14] One must exclude, however, costs that are solely attributable to the research component of the experiment. If, for example, the program incurs additional costs for outreach and intake in order to recruit a control group, those added costs should be netted out.

Treatment-Control Service Differential — Partial-Service Counterfactual

EXHIBIT 1

Group	Experimental Training (hours)	Nonexperimental Training (hours)	Total Hours of Training
Treatment Group	100	25	125
Control Group	0	50	50
Treatment-Control Difference	100	-25	75

overlooked. We will discuss the role of benefit-cost analysis in social experiments in a subsequent paper.

Interpreting the Treatment-Control Service Differential

The simplest type of experiment involves random assignment of program applicants to two groups: a treatment group that is allowed to enter the program and a control group that is not.[15] Treatment group members are allowed to receive all experimental program services, as well as any services outside the experiment for which they would otherwise have been eligible. Controls are excluded from the experimental program, but are otherwise free to do anything they wish, including receiving similar services from sources other than the experimental program. In this basic experimental design, the experience of the control group is intended to represent what would have happened to participants in the absence of the program—which we term the **counterfactual**. The treatment-control difference in outcomes measures the impacts of the program tested, *relative to the policy environment faced by the control group*. Interpretation of the impact estimates therefore requires an understanding of the policy environment faced by controls, as well as the treatment received by the treatment group.

It is, of course, critical to ensure that the policy environment faced by the controls is in fact the desired counterfactual. Therefore, we must also consider the relationship between the difference in policy environments faced by treatment and control group members and the policy question the experiment is intended to inform.

The no-service counterfactual

In the simplest case, there are no services or benefits outside the experimental program similar to those offered to the treatment group. Thus, the experimental contrast is between a treatment group that receives the experimental services and a control group that receives no similar services. Suppose, for example, that an experimental training program provides 100 hours of training to the treatment group. Controls are excluded from the experimental program; if no nonexperimental training is available, controls will receive zero hours of training, and the treatment-control service differential will be 100 hours. *Because the treatment-control difference in services received is identical to the full experimental treatment, treatment-control differences in outcomes can be interpreted as the full effect of the experimental treatment.*

The partial-service counterfactual

When services or benefits similar to those offered by the experiment are available outside the experimental program, the experimental services may *displace* some of the nonexperimental services that would have been received by the treatment group in the absence of the experiment. The treatment-control service differential, then, will be the net of the additional experimental services received by the treatment group and the reduction in the nonexperimental services they receive.

Exhibit 1 illustrates this situation for a hypothetical experimental training program. Suppose that, as before, the experimental program provides 100 hours of training to the treatment group. But now suppose that nonexperimental training is available in the community from existing programs (*e.g.*, the Employment Service or community colleges) and that, while controls are excluded from the experimental program, neither treatment nor control group members

[15] For simplicity of exposition, the discussion in this paper is framed in terms of randomly assigned individuals as the units of observation and analysis. In a subsequent paper, we will discuss random assignment of groups of individuals--e.g., classes, caseloads, or communities.

are prohibited from receiving nonexperimental services.[16] In this illustration, the average treatment group member receives 100 hours of training in the experimental program and 25 hours of nonexperimental training, while the control group receives an average of 50 hours of nonexperimental training. Thus, the experimental program displaces 25 hours of nonexperimental training in the treatment group, and the overall treatment-control service differential is 75 hours, rather than the full 100 hours of service provided by the experimental program. In this case, then, treatment-control differences in outcomes do not measure the full effect of the experimental treatment; they measure the **incremental impact** of an additional 75 hours of training.[17]

Relating Experimental Impacts to Policy Decisions

In the case of the no-service counterfactual, where there are no similar nonexperimental services available, there is a clear correspondence between the experimental impact estimates and a policy decision with respect to the program. Adopting the experimental treatment as an ongoing program will result in a net increase in services equal to the amount of service provided by the program.[18] The experimental impact estimates, which measure the full effect of the program in this case, therefore represent the effects that could be expected if the program were adopted.

The relationship of the impact estimates to policy decisions is less straightforward in the case of the partial-service counterfactual. If adoption of the program would displace existing services to the same degree that the experimental treatment did, then the service differential created by the experiment (75 hours of training in our example) will be a good measure of the service increment that would be created by an ongoing program, and the experimental impact estimates will be good estimates of the effects of adopting the program. However, there are several reasons why the experimental service differential may *not* be a good measure of the incremental services provided by an ongoing program.

First, the experiment may cause control group members to receive a different level of services than they would have in the absence of the experiment. For example, the outreach and recruiting activities of the experimental program may prompt some individuals who would not have sought services in the absence of the experiment to do so in its presence. When such individuals are assigned to the control group, they may seek out nonexperimental services that they would not have received in the absence of the experiment. Staff of the experimental program may also assist controls in finding nonexperimental services, as a "consolation prize" for being denied experimental services. Alternatively, exclusion from the experimental program could discourage controls from seeking out sources of help that they would have found in the absence of the experiment. All of these sources of **control group contamination** can cause the control level of services to differ from what it would have been in the absence of the experiment and, therefore, cause the treatment-control service differential to be different from that which would be created by adopting the program.

Second, adoption of the program may have **general equilibrium effects** that cause the resulting service increment to differ from the experimental treatment-control service differential. Suppose, for example, that adoption of the experimental treatment as an ongoing program would cause legislators to reduce funding for other programs providing similar services. In that case, the experimental service differential would not accurately represent the service increment that could be expected from adoption of the program. In the extreme case, where the new program is entirely funded by transferring resources from other programs, there would be no increase in services in the aggregate, just a relabeling and reallocation of services among the eligible population.[19]

In a subsequent paper, we will discuss the steps that can be taken in implementing the experiment to protect against control group contamination. Regardless of the precautions taken, however, in the end one can never be certain that this risk has been entirely avoided. Moreover, general equilibrium effects are, almost by definition, virtually

---

[16] For both ethical and logistical reasons, it is usually impossible to exclude either group from receiving existing nonexperimental services. As we shall see, receipt of some existing services may also be the relevant counterfactual for policy purposes.

[17] For illustrative purposes, this example treats experimental and nonexperimental training as interchangeable. Strictly speaking, the treatment-control differences in outcomes measure the effects of receiving 100 hours of experimental training and 25 hours of nonexperimental training vs. receiving 50 hours of nonexperimental training. Only if an hour of experimental training and an hour of nonexperimental training can be assumed to have the same effects can we net out the 50 hours of training received by the control group against the 125 total hours of training received by the treatment group and attribute the treatment-control difference in outcomes to the 75 hour difference.

[18] This statement assumes that all treatment group members participate in the program. We discuss below the case where some individuals assigned to the treatment group do not participate in the program.

[19] In this case, the appropriate counterfactual would be the program the experimental treatment would displace, not the *status quo* mix and level of nonexperimental services.

impossible to predict.[20] Therefore, caution must be exercised in the interpretation of the experimental impact estimates in cases where similar services are available outside the experiment.

Even if one cannot confidently assert that the treatment-control service differential represents the service increment that would result from adoption of the program, the impact estimates based on that differential may still be quite useful for policy. While those estimates may not correspond neatly to a policy action, such as adopting the program, they *do* provide valid estimates of the effects of a well-specified policy change—increasing the level of service by the amount of the treatment-control service differential.

In our training program example, for instance, we may not be able to say that adoption of this program would lead to an increase of 75 hours of training per trainee. But we *could* say that *if* an additional 75 hours of training of this type were provided, it would have the effects estimated by the experiment. Those effects could then be compared to the costs of providing 75 additional hours of training, in order to decide whether it would be worthwhile to provide that level of additional services.[21] Even if that level is not the level that would ultimately be provided through the policy process, such an analysis would provide a valuable benchmark for the likely social value of the program.

In practice, this may be the best that any type of study, experimental or nonexperimental, can hope to achieve in predicting the effects of a new program. It will nearly always be impossible to predict the exact form that the final version of a piece of social legislation will take. Even legislation patterned explicitly on a successful demonstration is likely to depart significantly from the demonstration intervention, as a result of the numerous other forces that impinge on the policy process. With an experiment, at least one can be confident that the impact estimates derived from the demonstration are unbiased measures of the effects of the service increment created by the demonstration.

Interpreting Treatment-Control Differences in Outcomes

The fundamental rationale of social experiments is that random assignment creates two or more groups of individuals who do not differ systematically in any way except the experimental treatment(s). Thus, any subsequent differences in their behavior that exceed the bounds of sampling error can confidently be attributed to the experimental treatment. In this section, we discuss the interpretation of those differences in outcomes (including what we mean by "the bounds of sampling error").

Admissible Comparisons

It is important to recognize that random assignment creates comparability between the *entire* treatment group and the *entire* control group; this fundamental strength of the experimental method *does not necessarily apply to subgroups of the treatment and control groups.*

Suppose, for example, that some of those assigned to the treatment group fail to participate in the experimental program. One cannot simply drop them from the sample and compare the outcomes of the program participants with those of the controls. To do so would inject into the analysis the very selection bias that experiments are intended to avoid, because treatment group members who chose to participate may well be systematically different from those who do not. Since there is no way to identify and exclude from the analysis the nonparticipants' counterparts in the control group, dropping the nonparticipants in the treatment group from the sample would create a fundamental mismatch between the two groups that could bias the impact estimates. (We discuss below how impacts on the subgroup who participate can be estimated in certain circumstances, but even that method requires that we first estimate the impact on all individuals randomly assigned.)

More generally, *it is not possible to derive experimental estimates of the impact of the treatment on* **"endogenously defined"** *subgroups*. By that we mean groups defined on the basis of events or actions that occur after random assignment. Because such events or actions may be affected by the experimental treatment to which the individual was assigned, they may define subgroups of the treatment and control group that are not comparable. Or, as in the case of program participation, the event may be applicable only to one group or the other; in such cases, there is no way even to identify the corresponding subgroup in the other experimental group.

This means that it is sometimes not possible to estimate experimentally program impacts on subgroups in which there is strong policy interest. For example, policy makers are often interested in whether impact varies with treatment "dosage"; it is frequently suggested that this question can be analyzed by comparing impacts on those who leave the program early with those who stay in the program longer. Because the behavior that determines length of stay oc-

[20] This same limitation applies to any nonexperimental analysis.

[21] In a subsequent paper we will discuss the use of impact estimates in a benefit–cost analysis.

curs after random assignment, one cannot analyze this issue experimentally. To compare self-selected (or program-selected) groups who received different levels of treatment would be analogous to studying the effects of medical care by comparing the health status of individuals who had short hospital stays with the health status of those who had long hospital stays.

Another common question that cannot be answered experimentally is, what were the impacts of the treatment after participants left the program, as distinct from those impacts that occurred while they were in the program? Because length of stay in the program typically varies within the treatment group and the concept "left the program" is not defined for the control group, there is no way to construct comparable time periods for treatment and control group members for analysis of this question.

While these restrictions on the analysis of experimental subgroups may seem severe, one can often construct an experimental comparison—either *ex ante* or *ex post*—that either answers the question or provides an acceptable substitute. If there is strong interest in the effects of alternative treatment dosages, for example, the experiment can be designed to answer that question, by randomly assigning individuals to alternative levels of treatment. This will create two treatment groups that are comparable to one another and to the control group.

And while post-program impacts cannot be precisely isolated, one *can* estimate impacts in each month, quarter, or year after random assignment, since time since random assignment is well-defined for both the treatment and control groups and cannot be affected by experimental status.[22] This allows one to estimate the impact of the program in the period when most, or all, of the participants have left the program.

One can also learn a great deal by analyzing subgroups that are *not* endogenously defined. In general, it is permissible to compare subgroups defined on the basis of events that occur, or characteristics that are measured, prior to random assignment. By definition, such events and characteristics cannot be affected by experimental status which, under random assignment, is uncorrelated with *all* preexisting characteristics; moreover, such characteristics are well-defined for both the treatment and control groups. Thus, for example, the difference in mean outcomes between women in the treatment group and women in the control group is a valid measure of the impact of the program on women.

[22] This analytic approach will be discussed in more detail in a subsequent paper.

It is often of great interest to estimate impacts for subgroups formed on the basis of demographic characteristics and baseline (*i.e.*, pre-random assignment) values of the outcomes of interest. For example, suppose we are estimating the impacts of a training program on the earnings and welfare benefits of AFDC recipients. It would be useful to estimate impacts for subgroups defined on the basis of age, education, ethnicity, length of time on welfare, or prior earnings or benefit level of the participant. This information would be useful in targeting the program on those recipients who could benefit most from it. And by identifying those recipients who were not benefiting from the program, it might also suggest ways to improve the program or at least target improvement efforts on the portion of the participant population where they are most needed.

Protecting Against Chance Differences in Outcomes Between Treatment and Control Groups

Random assignment guarantees that the only *systematic* difference between the treatment and control group is access to the experimental treatment. This means that if one were to replicate the experiment many times, on average the difference in outcomes between the treatment and control groups would equal the true impact of access to the program. We define the **expected value** of an estimator as its average value over many replications. When the expected value of the estimator equals the true value of the parameter it estimates, the estimator is said to be **unbiased**. Experimental treatment-control differences are unbiased estimators of the true effect of the experimental treatment on the (entire) treatment group.

In practice, of course, experiments are generally performed only once, not many times. Thus, while the *expected* value of the treatment-control difference in outcomes equals the true impact of the experimental treatment, in any one application it may differ from that value due to **sampling error**—chance differences between the two groups that result when specific individuals are randomly assigned to each group in a particular replication. Fortunately, statistical procedures can be used to place bounds on the size of the difference that could reasonably be attributed to sampling error and, therefore, to determine whether the observed treatment-control difference is likely to reflect more than sampling error.

Suppose, for example, that we randomly assign students either to go into a remedial education program or into a control group that receives no remediation. One year later, we compute the difference in grade point averages (GPA) between the two groups and find that the treatment group's grades are, on average, 0.6 points higher than the control

group's. Can we be sure that the program caused this difference? In addition to the effects of the program, grades will differ among students for any number of reasons that have nothing to do with the experimental program—*e.g.*, because of differences in native ability, motivation, health, or whether the student responds well to a particular teacher's pedagogical style.

It could be that, by the luck of the draw, more highly motivated students were randomly assigned to the treatment group than to the control group. If so, the treatment group's average GPA will be higher for this reason alone, and the treatment-control difference in grades will overstate the true impact of the intervention. Conversely, if those assigned to the treatment group were, on average, *less* motivated than the controls, the treatment-control difference will *under*state the true impact of the program.

How do we protect against mistakenly attributing these chance differences between the two groups to the program? The short answer is that *we use information about the natural variation in the outcome variable (in this case, GPA) across the sample to estimate the probability that a difference as large as that observed could occur entirely by chance.* To understand how this is done, we must first review some basic statistical concepts.

Statistically, the variation of an outcome across individuals is measured by its **variance**. The variance of an outcome Y is defined in terms of the deviation of particular values of Y (*e.g.*, each individual student's GPA) from the average value of Y in the population (which we denote μ_Y). Specifically:

$$\textbf{1} \qquad V_Y = E\left[(Y - \mu_Y)^2 \right]$$

where E[...] denotes expected value. Thus, the variance is the average of the squared deviations of Y around its mean.

The variance of Y measures how much individual Y values can be expected to vary around their mean. Now suppose we draw a sample of individuals and compute their mean Y. The mean of Y can be expected to vary less from one sample to another than Y does from one individual to another, because in the averaging process unusually high values and unusually low values offset each other. The variance of the *sample mean of Y* (denoted $\overline{Y}$) is the variance of Y divided by the number of observations in the sample, n:

$$\textbf{2} \qquad V_{\overline{Y}} = \frac{V_Y}{n}$$

Thus, the larger the sample, the less variable its mean will be.

The experimental impact estimate (I) is the difference between two means, the mean outcome of the treatment group (Y^T) and the mean outcome of the control group (Y^C):

$$\textbf{3} \qquad I = \overline{Y}^T - \overline{Y}^C$$

The variance of the difference between two independent means is the sum of their variances;[23] thus, the variance of the experimental estimator is given by the sum of the variances of the treatment and control group means:

$$\textbf{4} \qquad V_I = \frac{V_Y}{n_T} + \frac{V_Y}{n_C}$$

where n_T and n_C are the sample sizes of the treatment and control groups, respectively.[24]

This variance measures how much the experimental estimator would vary in repeated replications. As noted above, if the experiment were repeated a large number of times, each replication would yield a somewhat different estimate of program impact, because different sets of individuals would be assigned to the treatment and control groups in each replication. Taken together, the experimental estimates from many trials would form a pattern known as the **sampling distribution** of the experimental estimator. Exhibit 2 shows such a distribution (see next page). The height of the curve at any point along the horizontal axis in Exhibit 2 represents the proportion of trials that will yield impact estimates with that value. The area under the curve within any interval along the horizontal axis measures the proportion of trials that would yield estimates within that range. This area may therefore be interpreted as the *probability* that, in a *given* replication of the experiment, the experimental impact estimate will fall within that interval. For example, the shaded area in Exhibit 2 measures the probability that, in a given replication, the estimate of program impact would be greater than I_0.

Because experimental estimates are unbiased (*i.e.*, their average over many applications equals the true impact), we know that the sampling distribution of the experimen-

[23] Y^T and Y^C are statistically independent because they are based on two separate samples.

[24] For simplicity of exposition, we assume that the variance of Y in the treatment group equals the variance of Y in the control group. This need not be the case if the treatment affects the variance of the outcomes. If the variances differ, the formula would change slightly, but the principal conclusions presented here would remain the same.

tal estimator is centered on I^*, the true impact. Its shape is that of a normal (bell-shaped) curve with variance equal to the variance of the experimental estimator. As can be seen from equation 4 above, this variance depends on the variance of the outcome (V_Y) and the size of the experimental samples (n_T and n_C). The variance of the sampling distribution determines its overall shape—*i.e.*, how flat or peaked it is. The more highly variable the outcome Y is, the more widely will the sampling distribution be spread out along the horizontal axis. For any given variance of Y, the larger the sample sizes the more tightly clustered around I^* the distribution will be.

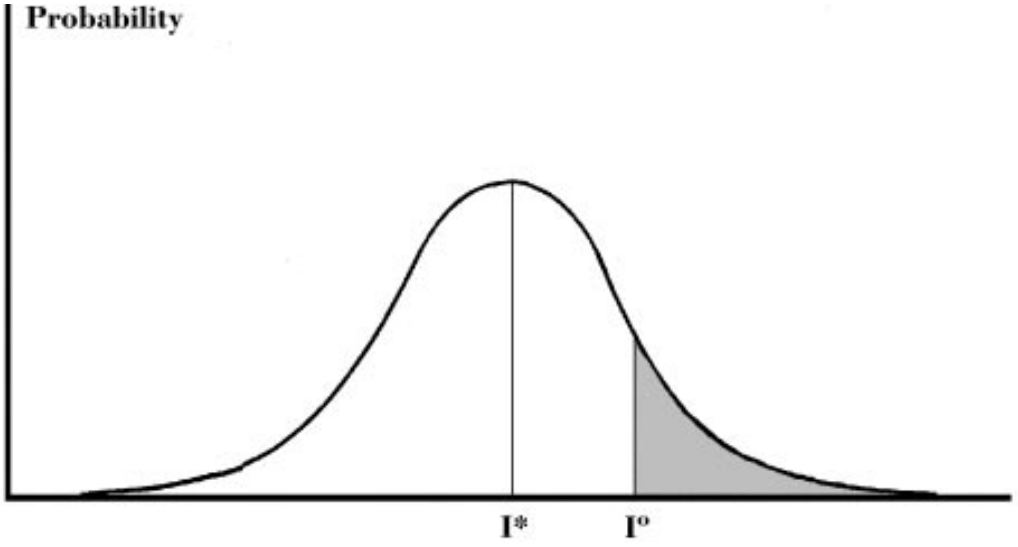
With the concept of the sampling distribution in hand, we can now return to our hypothetical remedial education experiment and ask the question, "How likely is it that we would obtain a treatment-control difference of 0.6 grade points or more by chance alone, when the true impact is zero?" To answer this question, we construct the sampling distribution that we would expect if the true impact of the program were zero. This distribution is centered on zero (the assumed true effect), with variance equal to the variance of the experimental impact estimate. The probability that an estimate of 0.6 or greater could have occurred by chance alone when the true impact is zero is given by the area under this distribution to the right of 0.6. We call this probability the **p-value** of the estimate. For example, a p-value of 0.15 would mean that, if the true impact is zero, we could expect an experimental impact estimate at least as large as 0.6 by chance alone 15 percent of the time.

The p-value can be used to *test the hypothesis that the true impact is less than or equal to zero—i.e.*, that the experimental program did not have a positive effect on Y. We term this the **null hypothesis**; the **alternative hypothesis** is that the true impact is positive. In such a test, we reject the null hypothesis of no positive effect and accept the alternative hypothesis of a positive effect if the probability that a treatment-control difference at least as large as that observed in the experiment could have occurred by chance alone (the p-value) is less than some pre-specified **significance level**. If the p-value exceeds that level, we cannot reject the null hypothesis. The significance levels usually used for this purpose are either 5 percent or 10 percent. That is, we require that the probability of obtaining an estimate as large as the observed result when the program truly has no positive effect be less than 1 in 20 (the 5 percent level) or 1 in 10 (the 10 percent level) before we accept the alternative hypothesis that the true impact is positive. Estimates that satisfy this criterion are said to be **statistically significantly greater than zero**.

Suppose, for example, that in the case of our hypothetical remedial education program the area under the sampling distribution to the right of 0.6 (the p-value of our experimental impact estimate) is 0.07. This means that 7 times out of 100 an estimate of this size or larger would be produced by chance alone if there were no true effect. If the significance level we have chosen for the test is 10 percent, we would reject the null hypothesis that the true impact is zero because—if the impact were really zero—

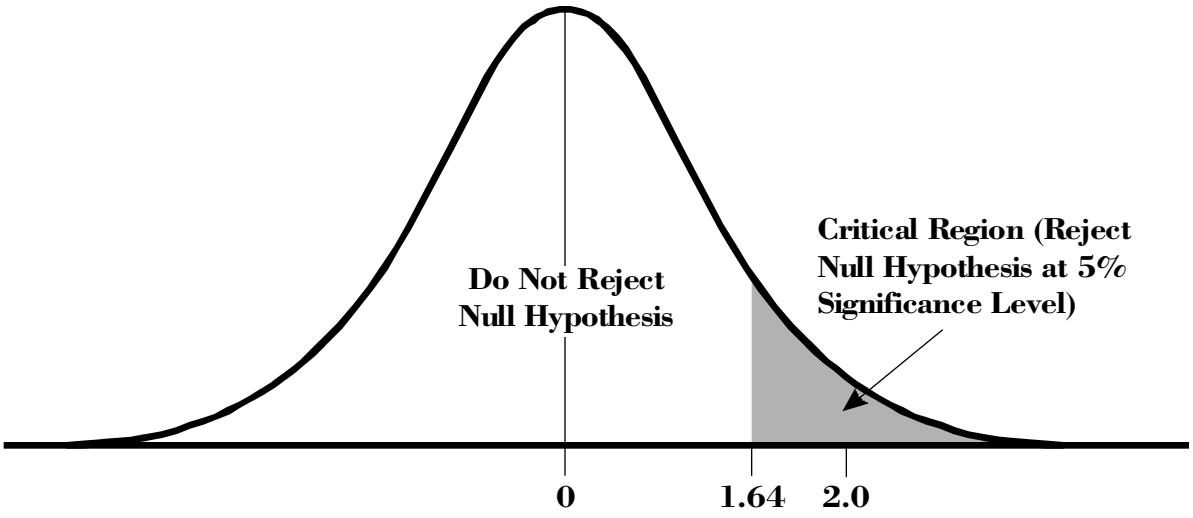
Sampling Distribution of the Experimental Estimator

EXHIBIT 2



Sampling Distribution of the *t*-Statistic

EXHIBIT 3



there is less than a 10 percent probability that an estimate as large as the one obtained would occur. Under the more stringent 5 percent significance level, however, we could not reject the null hypothesis and would have to entertain the possibility that the estimate differs from zero only because of sampling error.[25]

An equivalent test of the null hypothesis—and, in fact, the one that is usually used—is based on the ***t*-statistic**. The t-statistic is the impact estimate (I) divided by the square root of its variance, which is called the **standard error of estimate** (SEE_I):

$$\textbf{5} \qquad t = \frac{I}{\sqrt{V_I}} = \frac{I}{SEE_I}$$

The *t*-statistic measures the magnitude of the impact estimate in standard error units, rather than the natural units of the outcome variable (*e.g.*, grade points). Thus, the sampling distribution of the *t*-statistic is the *same* for all outcome variables under the null hypothesis of no effect. This means that, regardless of the outcome variable, one can test hypotheses using the same distribution of *t*-values; this distribution is available in published tables.

Like the sampling distribution of the impact estimate, the sampling distribution of the *t*-statistic has the property that the area under the curve within a given range is the probability that the *t*-statistic of the experimental impact estimate will fall within that range in any given application of the experiment, if the true impact is zero.

To test the null hypothesis of no program impact, one first establishes a **critical value** of the *t*-statistic that corresponds to the significance level of the test. For the 5 percent significance level, for example, the critical value is the point on the horizontal axis beyond which lies 5 percent of the area under the curve under the null hypothesis. This value can be determined from published tables of the *t*-distribution. The region outside the critical value is called the **critical region** (see Exhibit 3). If the *t*-statistic falls in the critical region, we reject the null hypothesis of no program effect because the probability of obtaining a *t*-statistic that large (*i.e.*, an impact estimate that many standard errors from zero) by chance alone is less than our chosen significance level.

For example, for large samples 5 percent of the area under the sampling distribution of the t-statistic lies to the right of 1.64.[26] (An impact estimate 1.64 standard deviations above zero would produce a *t*-value of this magnitude.) Thus, the critical value of the *t*-statistic in a test for a positive impact at the 5 percent significance level is 1.64, and the critical region includes all values of *t* greater than 1.64. If the *t*-statistic of the experimental estimate is greater than

[25] Although the conventional practice is to apply tests of significance to the experimental estimate, an alternative approach is simply to compute the p-value as a measure of the likelihood of an estimate at least as large as that obtained when the true impact is zero. Tests of statistical significance have the advantages that they yield a clear-cut yes or no decision on whether the experimental program had a real effect and they force the researcher to establish a standard of evidence in advance. The advantage of p-values is that they provide a more continuous, fine-grained measure of the risk that the estimate reflects only sampling error.

[26] Because the variance of the impact estimate depends on sample size (see equation 4), the sampling distribution of the *t*-statistic also depends on sample size. For sample sizes greater than about 30, however, the effect of sample size on the *t*-distribution is negligible.

1.64, we reject the null hypothesis that the program had a zero or negative effect on the outcome and conclude that it had a positive impact.

Application of the *t*-test is illustrated in Exhibit 3. In this example, the *t*-statistic equals 2.0; *i.e.*, the impact estimate is twice its standard error. This means that the estimate lies two standard errors away from zero. As noted above, the critical value for a test at the 5 percent significance level is 1.64; thus, the impact estimate lies in the critical region, the shaded area to the right of the critical value. This means that the probability that an impact estimate this large would be observed when there is no true effect is less than 5 percent and we must reject the null hypothesis of zero impact at the 5 percent significance level.

Up to this point, we have considered only the possibility that the experimental impact estimate might be significantly different from zero *in the positive direction*. Thus, the critical region for the test was confined to the right-hand tail of the sampling distribution. Such a test is called a **one-tailed test**. One-tailed tests are appropriate when a finding of a negative impact would have the same policy implications as a finding of zero impact. Suppose, for example, that one is testing a new approach to dropout prevention. A finding that the program actually *encourages* students to drop out would have the same policy implication as a finding of no impact—in either case, the approach should not be adopted. In such cases, we need only distinguish positive impacts from non-positive impacts; a one-tailed test does this.[27]

[27] One can, of course, also construct a one-tailed test to distinguish *negative* impacts from zero or positive impacts.

In some cases, though, there will be policy interest in distinguishing among positive, negative, and zero impacts. In those cases, a **two-tailed test** is appropriate. In a two-tailed test, the critical region is divided (usually equally) between the two tails of the distribution. Exhibit 4 illustrates the critical region for a two-tailed test. If the *t*-statistic falls in the critical region in either tail, we reject the null hypothesis of no effect. It should be noted that in a two-tailed test the *sum* of the areas in the critical region in the two tails is equal to the significance level. For example, for a test at the 10 percent significance level, the area in the critical region in *each* tail would equal 5 percent. Thus, the two-tailed test is a more stringent test for positive outcomes than the one-tailed test at the same significance level (*i.e.*, it is less likely to reject the null hypothesis).

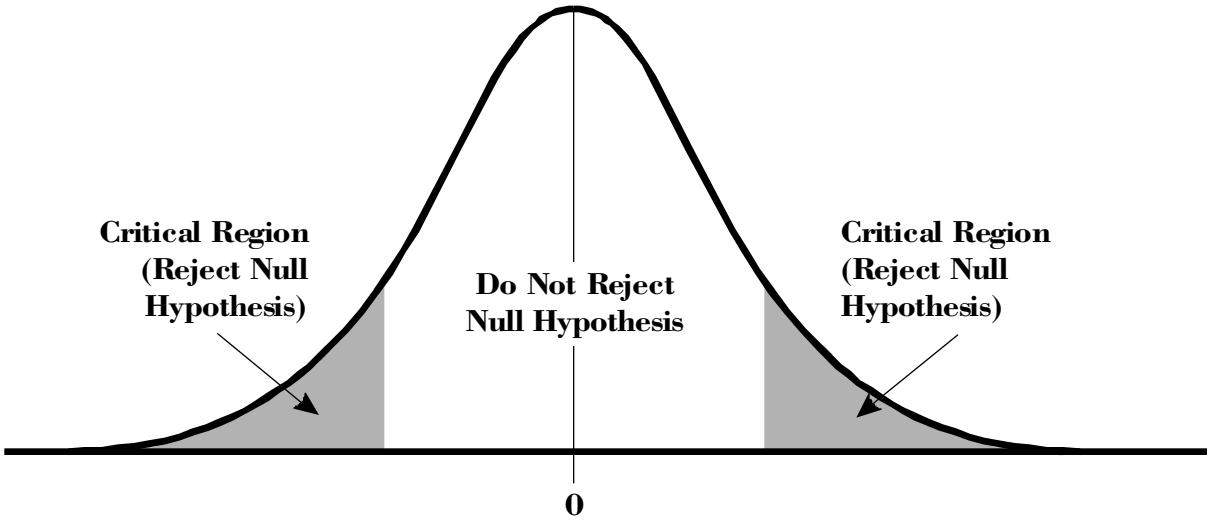
Interpreting the Results of Tests of Significance

In viewing the results of an experiment, it is important to understand what the results of tests of significance mean—and what they do not mean. A finding that the impact is significantly different from zero means that we can be reasonably sure that the experimental program had a nonzero impact. In such cases, there is only a small chance that the estimated impact would be as large as the one actually obtained if the true impact were zero.

The fact that the estimate is significantly different from zero does *not* mean, however, that we know the size of the true impact *exactly*. There is still sampling error attached to the estimate. The most we can say is that the true impact is likely to lie within a **confidence interval** around

Two-tailed *t*-Test

EXHIBIT 4



the estimate. The k percent confidence interval is the range around the experimental estimate that has a k percent chance of including the true impact. Confidence intervals are derived from the standard error of estimate, which measures the variability of the experimental impact estimate. The 95 percent confidence interval, for example, is the range 1.96 standard errors on either side of the experimental estimate. If the experiment were run repeatedly and this range calculated for each replication, 95 percent of those ranges would include the true impact.

Suppose, for example, that the estimated impact in our hypothetical remedial education program is 0.6, with a standard error of estimate of 0.2. Its t-statistic is 3.0 ($= 0.6 \div 0.2$), well above the critical value for statistical significance at the 5 percent level. So we can be reasonably sure that the true impact is greater than zero. But it may still be greater or less than 0.6. The standard error of estimate tells us that the probability is quite low—1 in 20 or less—that the true impact falls outside the range between 0.208 and 0.992 ($= 0.6 \pm 1.96 \times 0.2$).

Just as rejection of the null hypothesis of zero effect does not mean that we know the experimental impact exactly, *failure* to reject the null hypothesis does not mean that we know that the program had zero effect. Again, the estimate is subject to sampling error; the best we can do is to place a confidence interval around the estimate. Suppose, for instance, that in the example above the estimated impact of the remedial education program had been 0.2 grade points, rather than 0.6. With a standard error of 0.2, the 95 percent confidence interval around this estimate would have been the range from –0.192 to 0.592 ($= 0.2 \pm 1.96 \times 0.2$). Since this range includes both positive and negative numbers, we cannot be 95 percent confident that the program had a positive effect on grade point averages.

Another way of stating the same point is that, although a statistical test can lead one to *reject* the null hypothesis of zero effect, it can never lead one to *accept* the null hypothesis—*i.e.*, one never concludes that the true impact is *exactly* zero. One simply "fails to reject" the hypothesis that the true effect is zero. While this may seem like semantic hairsplitting, the difference between these two conclusions is enormous. The former says that we know the true effect *exactly*, whereas the latter says that we don't even know its direction! In policy terms, failure to reject the null hypothesis means that the evidence produced by the experiment was not strong enough to tell us whether the program tested was beneficial or not. This is far different from being convinced that it was not beneficial.

Even in this situation, however, the experiment may provide useful information about the size of the program effect. If the confidence interval around the estimated impact is so narrow that it includes only very small values on either side of zero, policy makers may not care whether the impact is positive or negative—even if there were a beneficial effect, it would have to be so small that it would not be sufficient to justify adopting the program. In contrast, if the confidence interval is so wide that it includes large beneficial values as well as zero effect, the program may still be worthwhile—the experiment has simply failed to prove that this is the case. In a subsequent paper, we will discuss ways to design the experiment to ensure that it has sufficient "power" to detect effects that are large enough to be important for policy, if they exist.

As this discussion suggests, statistical estimates—whether from an experiment or any other source—must be viewed in probabilistic terms. The true impact of a program can never be known with certainty. At best we can place it within some range with a high degree of confidence. While that may seem like a very limited objective, it is in fact an objective that is only attainable with experimental data. That is because experimental impact estimates are *known* to be unbiased and, therefore, depart from the true value *only* because of sampling error. Statistical tools are available to allow us to quantify the uncertainty attached to sampling error. Any nonexperimental estimation technique is subject not only to sampling error, but also to an unknown degree of bias. Thus, one cannot place *any* bounds on the true impact with nonexperimental methods unless one is prepared to assume that the estimate is unbiased.

Finally, it is important to recognize that even experimental estimates are valid only for the population from which the research sample was drawn and the treatment to which they were subjected. Thus, in assessing the evaluation results, it is important to consider whether the experimental treatment and population studied are the relevant ones for the policy issue at hand.

Inferring Impacts on Program Participants When Some Treatment Group Members Don't Participate

Random assignment ensures that the entire treatment group and the entire control group are comparable and that, therefore, the difference between their outcomes is an unbiased estimate of the average effect of the program on the treatment group as a whole. As pointed out earlier in this paper, this fundamental strength of the experimental method does not necessarily apply to subgroups defined by actions or events that occur after random assignment. A subgroup of

particular importance is those treatment group members who participate in the program. If not all treatment group members participate, the average effect of the program on the overall treatment group is likely to be "diluted" by the inclusion of nonparticipants on whom the program had little or no effect and, therefore, to understate the effect on participants. And policymakers are usually interested in the effect of the program on its participants, not on everyone who had the opportunity to participate.[28]

In social experiments, some degree of nonparticipation among treatment group members is almost unavoidable because program participation requires some action on the part of the sample member over which the experimenter has no control. In a typical experiment, individuals apply for services and go through an eligibility determination, and sometimes a needs assessment, before random assignment. Those found eligible and appropriate for the program are then randomly assigned and those assigned to the treatment group are informed that they are eligible to participate. Inevitably, some treatment group members fail to show up for services or decide at the last minute that they are no longer interested in the program. In a voluntary program, there is nothing that the program or the experimenter can do about such "no-shows".

Unfortunately, it is impossible to identify the control counterparts of the subgroup of treatment group members who participate, since controls do not have the option of participating in the program. And since participants are an endogenously defined subgroup, they will not necessarily be well-matched with the entire control group. Thus, one cannot obtain a direct experimental estimate of the impact of the program on participants. Fortunately, in some circumstances it is possible to infer this impact.[29]

To see how this can be done, we first express the average impact on the entire treatment group as a weighted average of the impact on participants and the average impact on nonparticipants, where the weights reflect the relative proportions of the two subgroups. Letting I represent the overall impact, I_p and I_n the impacts on participants and nonparticipants, and r_p and r_n the proportions of participants and nonparticipants in the treatment group, we have:

$$\mathbf{6} \qquad I = r_p I_p + r_n I_n$$

In the special case where the impact of the program on nonparticipants is zero ($I_n = 0$), the last term of this expression is zero and we have:

$$\mathbf{7} \qquad I = r_p I_p$$

Solving for I_p, we obtain:

$$\mathbf{8} \qquad I_p = \frac{I}{r_p}$$

That is, if the program had no effect on nonparticipants, the impact on participants is just the average impact on the overall treatment group divided by the proportion of the treatment group that participated (which we term the **participation rate**).[30] Dividing by the participation rate to obtain the impact on participants is known as the **no-show adjustment**.

Suppose, for example, that the estimated impact of an experimental training program on the average annual earnings of the entire treatment group is $1,000, but that only 80 percent of the treatment group participated in the program. The no-show-adjusted impact on the earnings of program participants would be $1,000 ÷ 0.80, or $1,250.

It is important to recognize that this derivation of the impact on participants makes no assumptions about the similarity or dissimilarity of participants and nonparticipants. The only assumption required is that the program has zero impact on nonparticipants. Under that assumption, the no-show adjustment will produce unbiased estimates of the impact on participants even when participants and nonparticipants are completely dissimilar and, therefore, participants are dissimilar to the control group. The adjustment simply averages the overall program effect across the participants, rather than across the entire treatment group.

In most voluntary programs, we can probably safely assume that the behavior of individuals who did not participate in the program at all was unaffected by the program. This assumption is not, however, valid in all circumstances. For example, in a mandatory work program for welfare recipients, some recipients might go to work

[28] While it may seem a forgone conclusion that the only policy interest should be in program participants, that is not necessarily the case. In some programs there may be policy interest in the average effects on the entire eligible population, as a measure of the program's effectiveness in addressing the broader problem that prompted interest in the program. For example, policymakers may want to know the effect of a training program for welfare recipients on the entire caseload, not just those who participate. In such cases, nonparticipation may be an important determinant of the program's effectiveness.

[29] The procedure described below is due to Bloom (1984).

[30] If we treat r_p as fixed (i.e., if we assume that it would be identical in repeated replications of the experiment), then the standard error of I_p is also $1/r_p$ times the standard error of I. Since both the estimate and its standard error have been multiplied by the same factor, the t-statistic of the estimated impact on participants is identical to the t-statistic of the estimated impact on the treatment group overall; tests of statistical significance on the two estimates therefore yield identical results.

and/or leave the welfare rolls in order to *avoid* participating in the program. It must be recognized that the no-show adjustment is only as valid as this underlying assumption and it should never be applied without careful consideration of the applicability of this assumption in the specific circumstances involved in the experiment.

A final point that should be recognized in applying the no-show adjustment is that the resulting impact estimates apply only to the participants in the experimental program. Nothing can be said about the impact the program would have had on nonparticipants had they participated; we simply did not observe the behavior they would have exhibited as program participants. Of course, if the same people would not have participated in an ongoing program, this is not a problem; in that case, the experimental participants represent the population of interest. This caveat is only important if the intake process in the experimental program is different from that which could be expected in a regular ongoing program, so that one could expect a different subset of those accepted into the program to participate in a regular program. The potential for obtaining an experimental participant group that is nonrepresentative of what one would expect in an ongoing program is a strong argument for making the experimental intake process as similar as possible to the intake process that would be used in an ongoing program. In a subsequent paper, we will discuss ways in which this can be done.

References

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review 8* (April): 225-46.

Botein, B. 1965. "The Manhattan Bail Project: Its impact in criminology and the criminal law process." *Texas Law Review* 43:319-31.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources 21* (Fall): 606-39.

Greenberg, David, and Mark Shroder. 1997. *Digest of the Social Experiments.* Washington, D.C.: Urban Institute Press.

Kershaw, David, and Jerilyn Fair. 1976. *The New Jersey Income-Maintenance Experiment. Volume I: Operations, Surveys, and Administration.* New York: Academic Press.

Newhouse, Joseph P. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, Mass.: Harvard University Press.

Wiseman, Michael. 1996. "State Welfare Reform Demonstration Projects." Tables presented at the 1996 Annual Workshop of the National Association of Welfare Research and Statistics (mimeo).